Filosofía e Inteligencia Artificial

Mauricio Enríquez Zamora

La inteligencia artificial (IA) es uno de los campos de estudio interdisciplinares más fascinantes en la actualidad, abarcando en sí ciencias tan diversas como la matemática, psicología, cibernética, informática y filosofía. Es sobre todo en su conexión con esta última que deseo en las siguientes líneas hacer una reflexión, es decir, en torno a los problemas de índole filosófica que entraña el estudio de este campo.

Tratando de asirnos a una definición más o menos aceptada de la IA, podemos afirmar que ésta consiste en la disciplina cuyo objetivo primordial es la comprensión, diseño y construcción de sistemas cuyo comportamiento pueda ser calificado de racional. Tal racionalidad no tendría que ser equivalente a la del hombre, aunque suele tomarse como referencia, lo cual nos permite distinguir los diferentes enfoques que existen al interior de esta disciplina, así como los distintos problemas filosóficos que le son inherentes.

Describiré primeramente estos distintos enfoques de la IA, poniendo de relieve la cuestión filosófica que subyace a cada uno de ellos. Adelanto aquí tales interrogantes: *¿Qué es la mente humana? ¿Cómo concebir actualmente la "racionalidad"? ¿Qué es la conciencia?* Cuestiones todas de naturaleza epistemológica, o que son competencia de esa rama de la filosofía llamada teoría del conocimiento. Pero, ya que la IA no se queda en la mera teoría sino que implica una práctica, una producción física, las repercusiones que esta puede tener en la vida de las personas nos conduce a problemas de tipo ético, jurídico o económico. Es preciso hacer también un somero análisis de estas cuestiones.

En último término, veremos que la reflexión en torno a estos problemas va al centro de la problemática antropológica expresada en la pregunta *¿Qué es el hombre?*; lo que nos obliga a considerar conceptos como los de *singularidad tecnológica* y *transhumanismo* con todas sus implicaciones filosóficas.

1. Enfoques de la inteligencia artificial.

1.1. El enfoque de la prueba de Turing.

En la IA se manifiestan cuatro maneras de concebir la inteligencia, de los cuales se derivan cuatro enfoques correspondientes de ella. Cada uno de estos enfoques o sigue como modelo al ser humano, o bien, a una concepción general de la racionalidad; además, cada enfoque atenderá primordialmente al puro pensamiento, o bien a la conducta y al pensamiento. Así, por ejemplo, la IA que atiende a la conducta y al modelo humano buscará comprender el comportamiento humano en conexión con sus motivaciones psicológicas. A este enfoque han aportado ciencias como la psicología y la etología. Asimismo buscará la construcción de un sistema que sea capaz de emular la conducta humana, en todas sus manifestaciones: emocional, conductual, cognitiva, etc.

La prueba propuesta por Turing en 1950 corresponde con esta forma de entender la inteligencia. Para el padre de la computación, la inteligencia artificial se evidencia suficientemente a través de una apariencia externa de humanidad. Si una persona puede llegar a confundir a una máquina con un ser humano, tal máquina puede recibir el calificativo de "inteligente". Turing enumeró las capacidades que dicha máquina debiera manifestar: procesamiento del lenguaje natural, representación del

conocimiento, razonamiento automático y aprendizaje automático. Pero si se quiere una manifestación de humanidad más completa habría que agregar otras dos habilidades: la visión computacional y la robótica.

La cuestión que subyace a este enfoque de la IA es la propia pregunta antropológica de *¿Qué es el hombre?* ¿Puede atribuirse a una máquina que manifiesta de manera efectiva todas las características distintivas del hombre (hasta el punto de no saber hallar la diferencia) el adjetivo de "humana"? Andrew, el robot que protagoniza la película *El hombre bicentenario*, aprendió a ser tan humano que exigió el derecho de portar este título. David, el niño robot del film *Inteligencia artificial*, fue programado para amar y necesitar de los otros y se convirtió en la única fuente de información acerca de la especie humana, dos mil años despúes, para los nuevos habitantes de nuestro planeta. ¿En qué sentido o en qué grado podemos decir que Andrew o David eran humanos?

Es un rasgo natural del ser humano la simpatía por los seres que se le asemejen en algún aspecto. ¿De dónde proviene nuestro afán por promover los "derechos" de los animales, si no es de una clase de indignación que nos produce ver la crueldad con que se les trata? Tenemos con ellos en común la capacidad de sentir. Así también podemos desarrollar muy fácilmente una empatía con esas máquinas humanizadas en un grado superlativo. Sin embargo, apelando al argumento de la *incapacidad*, podemos afirmar que no son humanos porque no pueden hacer ciertas cosas que sólo los humanos pueden hacer, como filosofar o angustiarse. No son humanos, sino que son como humanos. Este es el llamado sentido débil de la IA.

Curiosamente, la creación de IA nos orilla nuevamente a la pregunta antropológica, a replanteárrnosla. Definitivamente, la IA podría echar por tierra, por ejemplo, la caracterización aristotélica del ser humano como "ser racional". Las máquinas pueden ser tan racionales como nosotros, y aún con mejores resultados. Habrá que buscar, por tanto, otro aspecto que nos distinga.

<u>1.2. El enfoque del agente racional.</u>

Este es el enfoque que más se ha adoptado para la investigación, porque implica el de pensar racionalmente y supera a los de pensar y actuar como humanos. El sentido de esta superación estriba en que elimina aspectos concretos de lo humano que son de difícil control o cuyos fundamentos se desconocen. En cierto modo, esto significa sacarle la vuelta al problema de equiparar a la máquina con el hombre, centrando los esfuerzos simplemente en hacerla "racional". Aquí, la IA no es como el hombre, sino un agente racional.

¿Qué es un agente racional?

Un agente es algo que razona (agente viene del latín agere*, hacer). Pero de los agentes informáticos se espera que tengan otros atributos que los distingan de los "programas" convencionales, como que estén dotados de controles autónomos, que perciban su entorno, que persistan durante un periodo de tiempo prolongado, que se adapten a los cambios, y que sean capaces de alcanzar objetivos diferentes. Un agente racional es aquel que actúa con la intención de alcanzar el mejor resultado o, cuando hay incertidumbre, el mejor resultado esperado.*[1]

El agente racional es un ente que actúa en forma adecuada ante cierto tipo de problemas, de modo que sabe resolverlos. Sin duda que tal racionalidad es programada, pero se espera que el agente racional se mantenga con cierta autonomía, percibiendo su

1 Stuart, J. Russel; Peter Norvig.Inteligencia Artificial. Un enfoque moderno. p. 5.

entorno, representándose su conocimiento, razonando y aprendiendo a adaptarse a nuevas situaciones. La racionalidad que es alentada en este ente no es otra sino la racionalidad humana, pero incompleta, truncada. Le falta la conciencia.

Aunque la conducta de las máquinas inteligentes parece racional, esta racionalidad es inconciente. Apelando de nuevo al argumento de incapacidad diríamos que: "Una máquina es incapaz de saber de sí misma, de tener una identidad, deseos propios, etc." Pero entonces se reactiva en nosotros la cuestión del significado de nuestra propia conciencia, de nuestros deseos. Spinoza atribuía a todas las cosas un esfuerzo por perseverar en su ser (le denominó *conato*[2]), que en el caso humano es un esfuerzo conciente, pero nunca explicó el origen de esta conciencia. Sin embargo, hoy nos diría que las máquinas tienen también un conato propio, por el cual tienden a autoconservarse. En la medida en que una entidad dependa más de sí misma para existir, esta será más libre. Por tanto, los robots podrían categorizarse como libres, no tanto como el hombre, pero quizás más que los animales.

Spinoza fue uno de los pocos filósofos que ha defendido la idea de que el hombre no es algo colocado fuera de la naturaleza y que rompa sus leyes, sino que es parte de ella y sigue sus principios eternos, como cualquier otro fenómeno natural. En este sentido, Spinoza concibe al hombre como una especie de autómata, programado por Dios. Asimismo, empleamos medios naturales para existir, no sólo el aire, el agua y el alimento, sino también las prótesis u órganos artificiales. Nos naturalizamos, así como también humanizamos la naturaleza. ¿Quién puede negar que en esta dialéctica hombre-naturaleza, esta última pueda evolucionar hacia un nuevo tipo de humanidad?

2. Consecuencias prácticas de la IA.

Algunos de los problemas prácticos que resultan de la implementación de la IA son de índole social o jurídica, otros de tipo ético. Entre los del primer tipo tenemos que la automatización de la producción industrial acaba por desplazar al hombre en el trabajo. Esto es malo en sociedades donde los gobiernos no siguen una planeación de la economía que sirva al bienestar social, como debe ser. La causa del problema no es directamente la IA, sino la política que se implementa en tal sociedad. La robotización del trabajo en la industria debería significar la apertura de posibles actividades menos dañinas a la salud humana y más acordes con el desarrollo pleno del hombre.

Entre las consecuencias jurídico-políticas está la posibilidad de perder la privacidad:

La IA tiene el potencial de producir una vigilancia en grandes cantidades. Esta predicción puede convertirse en realidad: el sistema clasificado Echelon del gobierno americano "consiste en una red de envíos de escucha, campos de antenas y estaciones de radar; el sistema está respaldado por computadores que utilizan traducción de lenguajes, reconocimiento de voz y palabras clave que buscan pasar por la criba automáticamente todo el tráfico de llamadas telefónicas, correos electrónicos, faxes y telex".[3]

2 Spinoza, Baruch. Ética. Parte 3, proposición 6: "Cada cosa, en cuanto está en ella, se esfuerza por perseverar en su ser".

3 Stuart, J. Russel; Peter Norvig.Inteligencia Artificial. Un enfoque moderno. p. 1092.

Además de esta red de espionaje creada por gobiernos, no es menos perniciosa la de las empresas privadas: gigantes como Google, que utilizan IA en sus motores de búsqueda y traductores, también llegan a poseer demasiada información acerca de las personas, sin que éstas tengan un control de lo que se hace con ella.

Una consecuencia ética puede ser la pérdida de responsabilidad en lo que concierne a decisiones tomadas sobre la base del uso de máquinas inteligentes. Los sistemas expertos son programas que tienen una base de conocimiento de algún área específica (medicina, química, etc.) y que apoyan a los humanos expertos en la solución de problemas. Existen, por ejemplo, sistemas expertos que realizan diagnósticos médicos. La pregunta que surge sobre la responsabilidad, en caso de que halla error, es la siguiente: ¿la culpa es de la máquina o del ser humano?

Pero la consecuencia más radical de todas es la de que la IA puede significar el fin de la raza humana o de la raza humana tal como la conocemos. El primer término de esta disyunción implica que los seres humanos acabaremos siendo dominados, si no eliminados por las máquinas:

Casi cualquier tecnología tiene el potencial de hacer daño si se encuentra en las manos equivocadas, pero con la IA y la robótica, tenemos el problema nuevo de que las manos equivocadas podrían pertenecer a dicha tecnología. Incontables historias de ciencia ficción nos han alertado de los robots o de los "cyborgs" (robots-hombre) que se comportan de forma enajenada. Entre los primeros ejemplos se incluyen Frankenstein *(1818), de Mary Shelly... y la obra de Karel Capek* R.U.R. *(1921), en la que los robots conquistan el mundo. En cine, también tenemos* Terminator *(1984), en donde se combinan clichés de robots que conquistan el mundo con el viajar en el tiempo, y* Matrix *(1999), en donde se combinan robots que conquistan el mundo y el cerebro en una cubeta.*[4]

Este miedo a las máquinas inteligentes se puede explicar como se explica todo miedo a lo desconocido. En tanto permanecemos pasivos frente a las cosas, incluso ante las que son producto de nuestra actividad, tales cosas nos parecerán extrañas y hostiles; mas esto no pasa si con un entendimiento y una conducta activos buscamos explicarlas y llegamos a ver en ellas nuestra propia huella. Entonces nos identificamos con ellas y nos sentimos seguros.

El segundo término de la disyunción es un poco más objetivo. Se refiere a los cambios que pueden operarse en las máquinas o en el hombre mismo conforme avanza la tecnología. En ello está hipotéticamente implícito un momento que expresa un "salto cualitativo", es decir, un momento en que después de múltiples y progresivas acumulaciones de tecnología en el hombre o de humanidad en las máquinas, éstas lleguen a confundirse con nosotros. Los avances técnicos logrados hasta hoy nos hacen tangibles estas posibilidades.

El físico británico Stephen Hawking dijo en una entrevista para la BBC, el pasado 2014, que la inteligencia artificial augura el fin de la raza humana. Comentó que aunque mucha de esta tecnología ha sido hasta el momento de utilidad, teme que un desarrollo mayor podría hacer que ella misma se rediseñe y supere a la lenta evolución humana.

4 Ídem, p. 1093.

3. Singularidad tecnológica y transhumanismo.

Sobre todo en estas predicciones pesimistas del desarrollo de una IA amenazante a la existencia humana es donde se manifiesta una concepción fuerte, dura, de la IA. No se conciben a las máquinas inteligentes "como si fueran racionales", sino simple y llanamente "racionales", capaces de pensar por sí mismas, como lo hacemos las personas. Y señalan un momento hipotético en que aparece esta IA fuerte. Este momento se ha dado a llamar *singularidad tecnológica*. Claramente vinculado a este momento es el movimiento cultural del *Transhumanismo*:

Ray Kurzweil, en The age of Spiritual Machines *(2000) predice que hacia el año 2099 existirá "una fuerte tendencia hacia una fusión del pensamiento humano en el mundo de la inteligencia de la máquina que las especies humanas crearon inicialmente. Ya no existe una distinción clara entre los hombres y los computadores". Existe incluso una palabra nueva, trashumanismo, que se refiere al movimiento social real que ansía este futuro. Basta decir que estos temas presentan un reto para la mayoría de los teóricos que consideran la preservación de la vida humana y de las especies como algo bueno.*[5]

El transhumanismo, como movimiento social viene a respaldar ideológicamente entre el común de las personas lo que la ciencia y la tecnología pueden ofrecer a la evolución del hombre en un sentido positivo.

Bibliografía.

1. Spinoza, Baruch. Ética. Trotta. Barcelona. 2005.

2. Stuart, J. Russel; Peter Norvig. Inteligencia Artificial. Un enfoque moderno. Prentice Hall. Madrid. 2004.

5 Ídem, p. 1094.